Sprachtechnologie für Europa
Das vierte TELRI-Seminar in Bratislava

von Wolfgang Teubert

Vergangenen November trafen sich sechzig Sprachwissenschaftler, Computerlinguisten und Informatiker in Bratislava, der Hauptstadt der Slowakei im Herzen Europas. Sie kamen aus fast allen Ländern Europas, aus den USA und Usbekistan, aus Israel und sogar aus China, um darüber zu diskutieren, wie moderne Sprachtechnologie einen Beitrag zur Verständigung auch über Sprachgrenzen hinweg leisten kann. Das Thema dieses vierten TELRI-Seminars war Text Corpora and Multilingual Lexicography. Darum also geht es: Wenn uns der Computer beim Übersetzen eine wirkliche Hilfe sein soll, muss erst einmal geklärt sein, wie all die vieldeutigen Wörter mit ihren oftmals unscharfen Bedeutungen wirklich in Texten verwendet werden. Wir, die Sprachbenutzer, glauben das zu wissen, meistens wenigstens. Aber weiß es auch der Computer?

Solange sich Menschen direkt (oder über andere Menschen, beispielsweise Übersetzer und Dolmetscher) miteinander verständigen, gibt es in der Tat keine unlösbaren Probleme, wenigstens nicht für die europäischen Nationalsprachen. Natürlich müssen Wörterbücher gelegentlich aktualisiert werden. Letztlich jedoch können Übersetzer in allen Situationen, in denen sie nicht weiterwissen, ausreichend Hilfe in den vielseitigen Sammlungen von Sprachwissen finden, die die Linguistik bereitstellt.

Statt neue Fakten zu sammeln, begeben sich deshalb viele Sprachwissenschaftler –vorzugsweise auf die Suche nach Erklärungen für die als bekannt vorausgesetzten Sprachdaten – auf eine Reise, auf der sie sich gern von ihrer Introspektion, angewendet vorzugsweise auf die Modelle der Kognitionswissenschaften, leiten lassen. Es sind weniger die Unterschiede zwischen den Einzelsprachen als das allen Menschen gemeinsame Sprachvermögen, das die akademische Linguistik heute in seinen Bann zieht.

Doch Globalisierung und europäische Integration konfrontieren immer mehr Menschen mit polyglotten Situatio-

nen, denen sie auch dann noch hilflos gegenüberstehen, wenn sie die eine oder andere Fremdsprache gemeistert haben. Das ist schade. Denn eigentlich sollten wir Multilingualität ebenso wie kulturelle Vielfalt als Bereicherung erleben. Aber dazu würde gehören, dass wir uns mit diesen Sprachen direkt auseinander setzen können, bei allen Problemen, die das mit sich bringen mag. Oder sollte es wirklich dazu kommen, dass wir nur noch das zur Kenntnis nehmen, was uns in der neuen lingua franca Englisch vermittelt wird, und dass die anderen Nationalsprachen nur noch die Funktion von Geheimcodes haben, also dazu dienen, dem globalen Publikum Nachrichten vorzuenthalten? Wer auf der Welt liest schon deutsche, niederländische oder gar litauische Webseiten? Sollen wir uns also damit abfinden, dass alles, was nicht auf Englisch vorliegt, ungehört verhallt? Ein vielsprachiges Europa ist gottlob noch immer das Ziel europäischer Einigungsbestrebungen. Den mündigen Bürgern Europas muss es möglich gemacht werden, miteinander zu kommunizieren und ihre Bürgerrechte wahrzunehmen, ohne durch Sprachbarrieren daran gehindert zu werden. Dazu brauchen wir moderne Übersetzungshilfen, wirklich funktionierende elektronische Hilfsmittel und nicht Spielzeuge wie Taschenrechner mit eingebautem Wörterbuch. Nur so kann Europa vielsprachig bleiben und trotzdem demokratisch werden. Denn die polyglotten Eliten konnten sich schon immer verständigen: auf Latein, Französisch oder Englisch, je nach Jahrhundert. Doch bei aller Begeisterung für Europa haben es die letzten fünfzig Jahre nicht zuwege gebracht, Fremdsprachenkenntnisse wirklich zu verallgemeinern. Und selbst wer mehrere Fremdsprachen spricht, ist zumeist hilflos, wenn er oder sie sich für den Inhalt ungarischer oder slowenischer Internet-Seiten interessiert.

Alte Ansicht von Bratislava

Die multilinguale Sprachtechnologie versucht, den rapide wachsenden Bedarf nach Verständigung, nach Informationserschließung über Sprachgrenzen hinweg zu befriedigen. Bislang allerdings mit geringem Erfolg, wie jeder weiß, der sich schon einmal auf die von den Internetdiensten angebotene Übersetzungsoption verlassen wollte. Fünfzig Jahre Arbeit auf dem Feld automatischer Übersetzung haben außer einigen Nischenlösungen wenig Brauchbares gebracht, trotz in regelmäßigen Abständen abgegebener vollmundiger Versprechungen. Für Texte der Allgemeinsprache gibt es immer noch keine automatische Übersetzung. Woran liegt das?

Menschen können ihren gesunden Menschenverstand benutzen, um auch aus bruchstückhaften Daten die richtigen Schlussfolgerungen zu ziehen. Das geht Computern ab. Die multilinguale Sprachtechnologie ist in einem Flaschenhals steckengeblieben, weil dafür in unglaublich großem Umfang Daten und Fakten benötigt werden, also Sprachwissen, das bisher nicht in einer expliziten, aufbereiteten, vom Computer verarbeitbaren Weise zur Verfügung steht. Ein Übersetzer, der ständig französische Zeitungstexte liest, bekommt ein Gefühl dafür, wann er *Globalisierung* mit *globalisation*, wann mit *mondialisation* übersetzt. Für das automatische Übersetzungsprogramm zählen indessen nicht Gefühle, sondern Fakten: Im Kontext welcher Wörter findet man *globalisation*, im Kontext welcher Wörter benutzt man *mondialisation*? Fakten lassen sich prozedural verarbeiten, Gefühle nicht.

Sprachdaten lassen sich nur mit Empirie gewinnen. Deshalb heißt seit gut zehn Jahren das neue Zauberwort Sprachressourcen. Große elektronische Textsammlungen, so genannte Korpora, sind das Ausgangsmaterial, aus dem die benötigten Sprachdaten extrahiert werden müssen. Ein neuer Zweig der Sprachwissenschaft, die Korpuslinguistik, hat sich etabliert. Sie untersucht, welche Fakten wir sammeln müssen und wie sie sich, so vollautomatisch wie möglich, extrahieren lassen. Denn die erforderlichen Textkorpora, deren Größe inzwischen in Milliarden Wörtern gerechnet wird, lassen sich nicht mehr manuell (d. h. intellektuell) analysieren.

Das Institut für Deutsche Sprache ist die Heimat von TELRI, der Trans-European Language Resources Infrastructure. Das ist ein von der Europäischen Union finanziertes Langzeitvorhaben, an dem zur Zeit etwa vierzig Sprachinstitute in Europa (einschließlich der GUS) beteiligt sind. TELRI fördert den Aufbau von Sprachressourcen und leistet einen entscheidenden Beitrag zur korpuslinguistischen Forschung. Durch seine Kontakte zur Sprachindustrie weiß TELRI, welche Sprachdaten für die Entwicklung multilingualer Sprachtechnologie benötigt werden. Über TRACTOR, das TELRI Research Archive of Computational Tools and Resources, bedient TELRI die Bedürfnisse von Forschung und Entwicklung auf dem akademischen und dem industriellen Sektor. (Weitere Informationen bei www.telri.de und www.tractor.de)

Die TELRI-Seminare sind wichtige Foren, auf denen sich Forschung und Anwendung begegnen, neue Erkenntnisse ausgetauscht und innovative Projekte vereinbart werden. Hier werden die gemeinsamen Aktivitäten der TELRI-Partner der Fachöffentlichkeit präsentiert. Wie wichtig diese Seminare inzwischen geworden sind, zeigt die Teilnahme des für Europäische Angelegenheiten zuständigen Generaldirektors beim slowakischen Ministerpräsidenten, Igor Hajdusek, an unserem Seminar. Er trägt die Verant-

wortung dafür, dass zehntausende Seiten von Verträgen, Richtlinien und Vorschriften rechtzeitig zum EU-Beitritt ins Slowakische übersetzt werden, sowie umgekehrt ein Berg slowakischer Dokumente in die EU-Sprachen. Ohne DV-Unterstützung geht das nicht. Eines der Projekte, die in Bratislava diskutiert wurden, sieht denn auch den Aufbau einer multilingualen terminologischen Datenbank für den Wortschatz Recht und Verwaltung vor, mit den Sprachen der Beitrittsländer Litauisch, Polnisch und Slowakisch und den EU-Arbeitssprachen Französisch, Englisch und Deutsch. Dabei erfolgt die Verknüpfung zwischen den verschiedenen Sprachen über ein sogenanntes Parallelkorpus, das aus den relevanten EU-Dokumenten in den Versionen der sechs Einzelsprachen besteht. Parallelkorpora enthalten in nuce das Sprachwissen, das für die computergestützte Übersetzung benötigt wird. Gelingt es uns, dieses Wissen automatisch zu extrahieren, haben wir in ihnen den Ausweg aus der Dauerkrise gefunden, in der sich die automatische Übersetzung befindet. Trotz aller vorläufigen Erfolge steckt die Korpuslinguistik, wenn es um das Was und das Wie der automatischen Extraktion von Sprachwissen geht, noch in ihren Kinderschuhen. Das TELRI-Seminar in Bratislava hat gezeigt, dass es noch ein weiter und mühsamer, wenngleich erfolgversprechender Weg ist hin zu einer Übersetzungsplattform, die brauchbare Rohübersetzungen allgemeinsprachlicher Texte liefert.

Dr. Wolfgang Teubert ist wissenschaftlicher Mitarbeiter am Institut für Deutsche Sprache in Mannheim.